International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Credit Risk Model: Research on Credit Risk Categorization model using XGBoost

Puneet Singh[1], Shubha Mishra[2], Gargi Porwal[3], Prakhar Saxena[4], Rishabh Tripathi[5]

[1]Bachelor of Technology in Computer Science Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow
[2]Assistant Professor, Computer Science Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow
[3]Bachelor of Technology in Computer Science Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow
[4]Bachelor of Technology in Computer Science Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow
[5]Bachelor of Technology in Computer Science Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow

---

Abstract - Machine Learning is a subset of Artificial Intelligence technology that enables systems to learn and make decisions on their own. These systems can make accurate decisions by analyzing datasets and information without the need for explicit programming. This paper mainly introduces the application of machine learning algorithm (XGBoost) in credit risk assessment in the financial industry. Credit risk assessment is a significant challenge for banks to assess credit worthiness among many applicants and plays a very crucial role in the profitability of banks.

Our research paper addresses the limitations and complexities in the current models in the market that lack interpretability and transparency. The methodology section introduces the application of the random forest model in financial quantification, including the model principle, feature importance calculation and experimental design. Utilizing the dataset comprising of more than one lakh users from the CIBIL and other banks and then through the exploratory data analysis, feature selection and model construction of the credit risk prediction dataset, the construction and evaluation process of the CRM model is demonstrated.

Finally, the performance of the XGBoost model after hyperparameter tuning is evaluated and compared with other models to demonstrate its advantages and applicability in financial quantification.

Key Words: XGBoost, Machine Learning, Credit Risk Model, Random Forest

1.INTRODUCTION

In the financial system risk refers to the likelihood of a borrower defaulting in their obligations. It is a critical area of concern for institutions which is important to ensure informed lending, decisions, portfolio stability and profitability. However traditional methods like credit score and logistic regression have been increasingly challenged by the complexity of modern borrower profiles and dynamic nature of financial markets. The approach relies heavily on historical data and often fails to adapt emerging trends or hidden patterns.

Machine Learning had emerged as a transformative tool in financial analytics and decision making, leveraging its ability to process vast datasets and make accurate data-driven predictions. It helps in enhancing predictive data accuracy and automation of repetitive decision-making process this minimizing manual intervention.

The research is focused on proposing a robust credit risk model based on advanced machine learning techniques for overcoming weaknesses of conventional credit risk models

2. Body of Paper

1 MATERIALS AND METHODS

XGBoost is an optimized implementation of Gradient Boosting and is a type of ensemble learning method that is designed for efficiency, speed and high performance. It uses decision tree as its base leaners and combines them sequentially to improve model's performance. Now each subsequent tree is trained to correct the errors made by the previous tree.

The reason for using this algorithm is that XGBoost has built-in parallel processing to train models on large datasets quickly. It is also highly scalable and efficient to be able to handle large datasets. It also provides various regularization techniques that we will be using later in the tuning process.

1.1 Data Pre-processing: Data to be used has been collected from CIBIL and few other sources from the banking institution which comprises of more than 50 columns and over one lakh users. This data needs to be processed as it contains various inconsistencies. Firstly, null values are removed and in case of columns having a large percentage of values as NULL, we would remove the column itself. The initial preprocessing results in preparation of two main data- One from CIBIL and Second from other sources.

Fig - 1 shows the data from banking institutions and Fig - 2 is the dataset collected from the CIBIL sources.

```
Data columns (total 25 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   PROSPECTID           51196 non-null  int64
 1   Total_TL             51196 non-null  int64
 2   Tot_Closed_TL        51196 non-null  int64
 3   Tot_Active_TL        51196 non-null  int64
 4   Total_TL_opened_L6M  51196 non-null  int64
 5   Tot_TL_closed_L6M    51196 non-null  int64
 6   pct_tl_open_L6M      51196 non-null  float64
 7   pct_tl_closed_L6M    51196 non-null  float64
 8   pct_active_tl        51196 non-null  float64
 9   pct_closed_tl        51196 non-null  float64
 10  Total_TL_opened_L12M 51196 non-null  int64
 11  Tot_TL_closed_L12M   51196 non-null  int64
 12  pct_tl_open_L12M     51196 non-null  float64
 13  pct_tl_closed_L12M   51196 non-null  float64
 14  Tot_Missed_Pmnt      51196 non-null  int64
 15  Auto_TL              51196 non-null  int64
 16  CC_TL                51196 non-null  int64
 17  Consumer_TL          51196 non-null  int64
 18  Gold_TL              51196 non-null  int64
 19  Home_TL              51196 non-null  int64
 20  PL_TL                51196 non-null  int64
 21  Secured_TL           51196 non-null  int64
 22  Other_TL             51196 non-null  int64
 23  Age_Oldest_TL        51196 non-null  int64
 24  Age_Newest_TL        51196 non-null  int64
dtypes: float64(6), int64(19)
```

Fig - 1

```
Data columns (total 54 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   PROSPECTID                41966 non-null  int64
 1   time_since_recent_payment 41966 non-null  int64
 2   num_times_delinquent      41966 non-null  int64
 3   max_recent_level_of_deliq 41966 non-null  int64
 4   num_deliq_6mts            41966 non-null  int64
 5   num_deliq_12mts           41966 non-null  int64
 6   num_deliq_6_12mts         41966 non-null  int64
 7   num_times_30p_dpd         41966 non-null  int64
 8   num_times_60p_dpd         41966 non-null  int64
 9   num_std                   41966 non-null  int64
 10  num_std_6mts              41966 non-null  int64
 11  num_std_12mts             41966 non-null  int64
 12  num_sub                   41966 non-null  int64
 13  num_sub_6mts              41966 non-null  int64
 14  num_sub_12mts             41966 non-null  int64
 15  num_dbt                   41966 non-null  int64
 16  num_dbt_6mts              41966 non-null  int64
 17  num_dbt_12mts             41966 non-null  int64
 18  num_lss                   41966 non-null  int64
 19  num_lss_6mts              41966 non-null  int64
 20  num_lss_12mts             41966 non-null  int64
 21  recent_level_of_deliq     41966 non-null  int64
 22  tot_enq                   41966 non-null  int64
 23  CC_enq                    41966 non-null  int64
 24  CC_enq_L6m                41966 non-null  int64
 25  CC_enq_L12m               41966 non-null  int64
 26  PL_enq                    41966 non-null  int64
 27  PL_enq_L6m                41966 non-null  int64
 28  PL_enq_L12m               41966 non-null  int64
 29  time_since_recent_enq     41966 non-null  int64
 30  enq_L12m                  41966 non-null  int64
 31  enq_L6m                   41966 non-null  int64
 32  enq_L3m                   41966 non-null  int64
 33  MARITALSTATUS             41966 non-null  object
 34  EDUCATION                 41966 non-null  object
 35  AGE                       41966 non-null  int64
 36  GENDER                    41966 non-null  object
 37  NETMONTHLYINCOME          41966 non-null  int64
 38  Time_With_Curr_Empr       41966 non-null  int64
 39  pct_of_active_TLs_ever    41966 non-null  float64
 40  pct_opened_TLs_L6m_of_L12m 41966 non-null  float64
 41  pct_currentBal_all_TL     41966 non-null  float64
 42  CC_Flag                   41966 non-null  int64
 43  PL_Flag                   41966 non-null  int64
 44  pct_PL_enq_L6m_of_L12m    41966 non-null  float64
 45  pct_CC_enq_L6m_of_L12m    41966 non-null  float64
 46  pct_PL_enq_L6m_of_ever    41966 non-null  float64
 47  pct_CC_enq_L6m_of_ever    41966 non-null  float64
 48  HL_Flag                   41966 non-null  int64
 49  GL_Flag                   41966 non-null  int64
 50  last_prod_enq2            41966 non-null  object
 51  first_prod_enq2           41966 non-null  object
 52  Credit_Score              41966 non-null  int64
 53  Approved_Flag             41966 non-null  object
dtypes: float64(7), int64(41), object(6)
```

Fig - 2

1.2 Exploratory Data Analysis: After data cleaning we will merge the datasets using inner join and perform EDA to understand and extract information from the dataset. These insights will be helpful while processing and working with features. The figures below represent the distribution of data among various features like gender(Fig -6) , approved flag (Fig - 3), marital status (Fig - 5) and last prod_enq2 (Fig - 4).

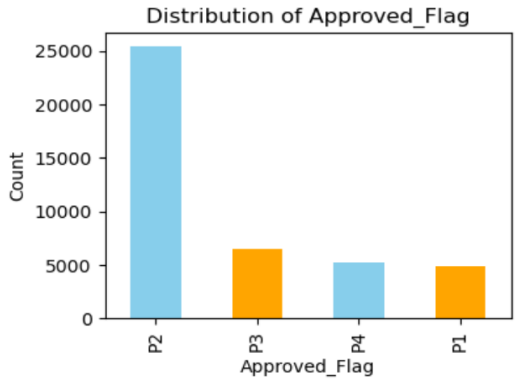

Fig - 3

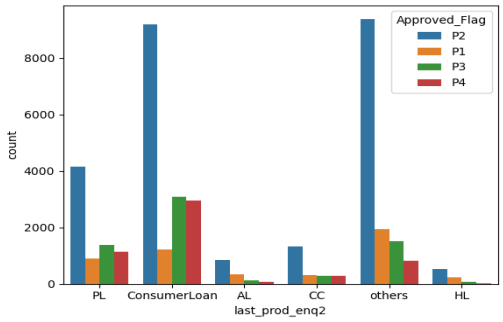

Fig -4

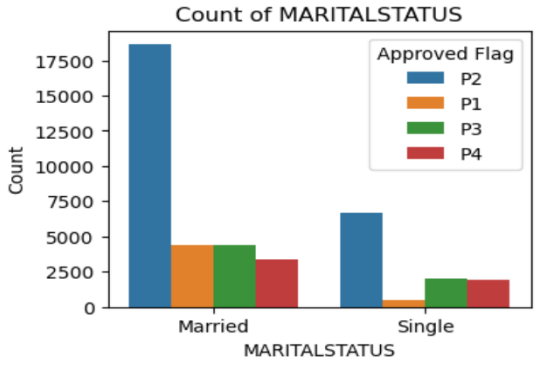

Fig - 5

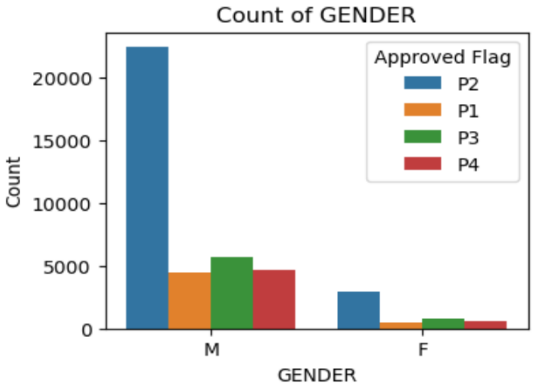

Fig - 6

1.3 Feature Selection: This article will focus on the key steps of feature selection and extraction that will be needed for building an efficient model. We will be using correlation analysis and feature coding to extract key information.

Firstly, we will perform CHI square test for the categorical columns to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table. Then for the numerical values we will perform VIF (Variance Inflation Factor) to check how correlated predicator variable is with other variables. For this specific model we have considered the variables with VIF values less than or equal to 6 which suits the idea of the model.

Next, we will perform Anova to compare the amount of variation between the group means to the amount of variation within each group. If the between-group variation is substantially larger than the within-group variation, it suggests that the group means are likely different. Fig - 7 shows the correlation of 'Age_Oldest_TL' with other columns and Fig - 8 represents the correlation matrix.'

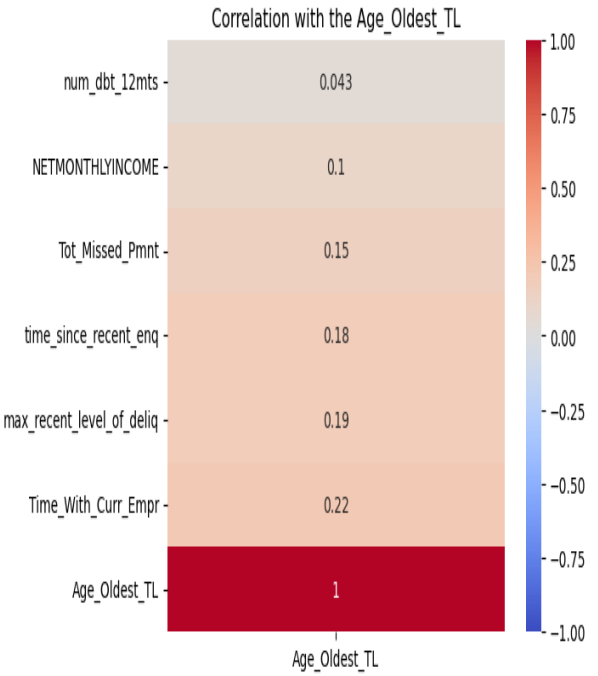

Fig - 7

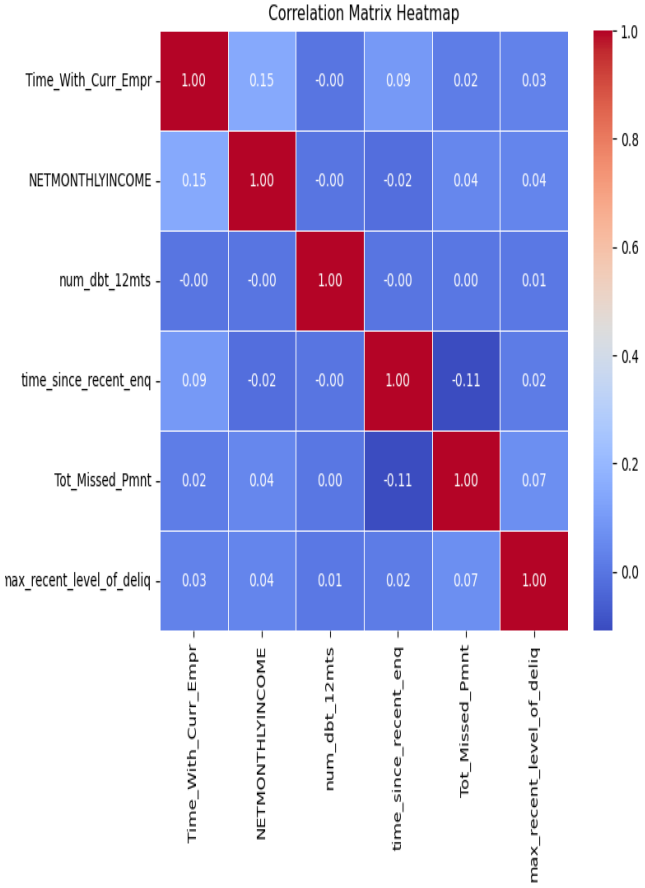

Fig - 8

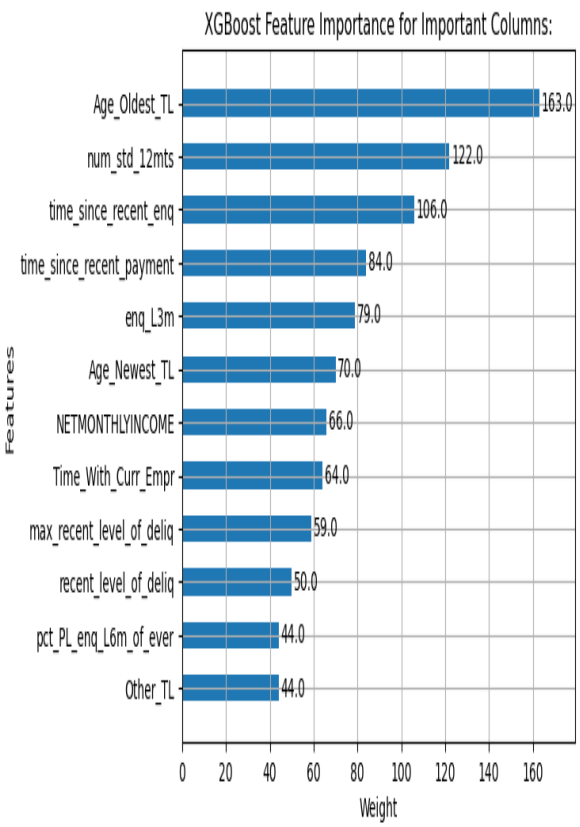

Fig - 10

1.4 Label Encoding: Label encoding is necessary as there are some categorical columns in the data. We will convert these categorical data into numerical so that they can be fitted by machine learning models which can only take numerical data. Fig - 9 shows the list of categorical columns present int the dataset.

```
List of Categorical columns :
MARITALSTATUS
EDUCATION
GENDER
last_prod_enq2
first_prod_enq2
Approved_Flag
```

Fig - 9

1.5 Model Building: Now different models are trained based on random forest, XGBoost and decision tree and their accuracy are checked. It is noted that XGBoost shows best results for this dataset among other algorithms. XGBoost shows the Test Accuracy of 0.77779 with the parameters 'learning_rate': 0.2, 'max_depth': 3, 'estimator's': 200.
Fig - 10 shows the XGBoost performance indicators for the important features.

1.6 Hyperparameter Tuning: To further increase accuracy Hyperparameter tuning is done to find the best set of hyperparameters for the CRM model. Hyperparameter tuning allows us to tweak model performance for optimal results. So we construct the parameter grid and then iterate the algorithm on each set of combinations in the grid and check for the optimal values of the parameter where the model gives better results.

After the iteration we get the optimal values for the parameters that are:
num_class=4,
colsample_bytree=0.9,
learning_rate=1,
max_depth=3,
alpha=10,
n_estimators=100
and the accuracy of the model increases to 0.8101.

2 RESULTS AND DISCUSSION

In the given data, the performance of the XGBoost algorithm as seen from the results on unseen data (Test data) is very satisfactory. According to the results as shown in Fig -11 XGBoost offers accuracy of 0.77 which after hyperparameter tuning increases up to 0.82 which is higher than both Random Forest and decision tree and also works well for p2 and p3 classes. The proposed model provides a comprehensive approach to credit risk analysis that leverages advances ML

techniques and incorporates features to meet both industry needs as well as regulatory needs.

The ability to handle unbalanced datasets and superior accuracy even in complex and tough situations is what makes XGBoost our choice to use for credit risk model building. XGBoost uses sequential ensemble learning with tree and is highly scalable and efficient. Therefore, in the credit risk assessment where a large dataset is processed and accuracy is important, XGBoost has been used and comparison with other similar models has been done to provide a better comprehension of the application of models in Fintech problems.

Table -1 clearly compares the accuracy of the models and their performance for the different classes where XGBoost is better in most cases compared to other models.

Table -1: Comparison of Models

Metric	Random Forest	Decision Tree	XGBoost
Accuracy	0.7604	0.7100	0.7700
Class p1 Precision	0.8247	0.7270	0.8121
Class p1 Recall	0.6989	0.7314	0.7782
Class p1 F1 Score	0.7566	0.7292	0.7948
Class p2 Precision	0.7929	0.8084	0.8191
Class p2 Recall	0.9263	0.8288	0.9086
Class p2 F1 Score	0.8545	0.8184	0.8615
Class p3 Precision	0.4269	0.3448	0.4511
Class p3 Recall	0.2033	0.3023	0.2857
Class p3 F1 Score	0.2755	0.3221	0.3498
Class p4 Precision	0.7249	0.6340	0.7429
Class p4 Recall	0.7185	0.6520	0.7234
Class p4 F1 Score	0.7216	0.6429	0.7330

3. CONCLUSIONS

To sum up, this paper clearly discusses the application of XGBoost in Credit risk model and the paper also proves the efficiency and effectiveness of model in prediction of credit risk. With the continuous development in AI technology and emergence of new technologies like big data and Blockchain there is always a need of further improvement to increase the model effectiveness with changing scenarios and applications. On the other hand, the risks and challenges are also increasing that must be faced with responsible use of the technologies. In the upcoming times we need to keep further exploring the ways of enhancing the security and the accuracy of the model. Through continuous innovation and improvement of artificial intelligence technology, we can keep this model updated and reliable even with the changing times because as financial markets evolve, models that can continuously learn, integrate diverse data sources and big data, and provide actionable insights will be indispensable.

REFERENCES

1. Uddin, M. S., Chi, G., Al Janabi, M. A. M., & Habib, T. (2023). Leveraging Random Forest in Micro-Enterprises Credit Risk Modeling for Accuracy and Interpretability.
2. Yu, C., Jin, Y., Xing, Q., Zhang, Y., Guo, S., & Meng, S. (2024). Advanced User Credit Risk Prediction Model using LightGBM, XGBoost and Tabnet with SMOTEENN. arXiv preprint, 2408.
3. Bielecki, T. R., & Rutkowski, M. (2021). Credit Risk: Modeling, Valuation, and Hedging. Springer-Verlag.
4. Doko, F., Kalajdziski, S., & Mishkovski, I. (2021). Credit Risk Model Based on Central Bank Credit Registry Data. Journal of Risk and Financial Management, 14(138). MDPI.
5. Cao, L., Yang, Q., & Yu, P. S. (2021). Data Science and AI in FinTech: An Overview. International Journal of Data Science and Analytics, 12(1), 81–99.